文章编号：2095-2163(2021)05-0193-05 中图分类号：TP391 文献标志码：A

基于 BiGRU-Capsule 的多标签文本分类

肖萍婉，王子牛，高建瓴
（贵州大学 大数据与信息工程学院，贵阳 550025）

摘 要：传统的文本分类一般采用单标签形式，但现实生活中多标签文本比单标签文本具有更广泛的应用场景。本文提出一种 BiGRU-Capsule 模型的多标签文本分类方法，该方法首先通过嵌入层将输入的文本序列转化为向量表示；然后通过 BiGRU 和 Capsule 提取文本特征；最后使用 sigmoid 分类器进行分类。为确保数据量足够，利用今日头条 2018 新闻标题多标签语料数据集进行实验，将胶囊网络模型作为对比模型进行多标签文本分类实验与分析。实验结果表明：本文模型的多标签文本分类效果得到有效提升。

关键词：多标签文本分类；BiGRU；Capsule

Multi-label text classification based on BiGRU-Capsule

XIAO Pingwan, WANG Ziniu, GAO Jianling
(School of big data and information engineering, Guizhou university, Guiyang 550025, China)

【Abstract】 Traditional text classification generally uses a single-label form. However, in real life, multi-label text has a wider range of application scenarios than single-label text. Different from single-label text classification, multi-label text classification means that a sample may correspond to one or more labels. In this paper, a model of BiGRU-Capsule is proposed. First, the text is converted into a vector representation through the embedding layer of the input text sequence; then the text features are extracted through BiGRU and Capsule; finally, the sigmoid classifier is used for classification. In order to ensure that the amount of data is sufficient, experiments were conducted on the multi-label corpus dataset of today's headlines 2018 news headlines. The capsule network model is used as a comparative model for multi-label text classification experiments and analysis. The final experimental results show that the multi-label text classification effect of this model has been effectively improved.

【Key words】 multi-label text classification; BiGRU; Capsule

0 引 言

在大数据时代，每天都在产生各种类型的数据，数据量大且具有多样性。多标签文本在日常生活中十分常见，例如：一条微博可能同时标注“明星”、“综艺”、“搞笑”、“娱乐”等多个标签；一则体育新闻可能同时标注“体操”、“奥运会”、“体育”等标签。多标签文本分类在现实生活中有许多实际应用，如视频注释、主题识别[1]、情感分析[2]、信息检索[3]等。因此，多标签文本分类任务是自然语言处理领域一个十分重要却又富有挑战性的任务。

1 相关研究

目前，多标签文本分类的研究方法可分为 3 种类型，分别是算法适应方法、问题转换方法和神经网络方法。算法适应方法是根据已存在的传统单标签文本分类算法，进行相对应的改进后，得到适应处理多标签分类的算法。Elissee 等人提出 Rank-SVM (Ranking Support Vector Machine)方法[4]，是将经典的支持向量机运用到多标签分类中；陆凯等人提出的 ML-KNN 方法[5]，是先利用 K 近邻算法得到近邻样本的标签，然后未知示例的标记集合是通过最大化后验概率推理得到。问题转换方法是将多标签分类任务转化为传统的单标签分类任务，目前单标签分类任务已经有许多成熟的算法可以选择。如，二元分类算法 BR(Binary Relevance)[6]，是将多标签学习问题分解为多个独立的二元分类问题，但存在缺乏发现标签间相互依赖的能力，这将会导致预测标签的性能降低；标签统一算法 LP(Label Powerset)[7]，是将每个有可能的标签重新整合成一组新的标签集合，再将问题转化为单标签分类任务；分类器链算法 CC(Classifier Chains)[8]，是将多标签学习任务转换为二元分类问题链，链上每个节点对应于一个标记，通过模拟标签之间的相关性进行分类；文献[9]设计了基于流式多目标回归器 iSOUP-Tree 的多标签分类方法；文献[10]设计了一种基于

作者简介：肖萍婉(1996-)，女，硕士研究生，主要研究方向：电子与通信工程；王子牛(1961-)，男，硕士，副教授，主要研究方向：计算机应用系统、信息系统；高建瓴(1969-)，女，硕士，副教授，主要研究方向：数据分析、数据库应用。

收稿日期：2021-01-10

去噪自编码器和矩阵分解联合嵌入多标签分类算法 Deep AE-MF;如此等等。上述两个类型的方法都是依赖于大量特征工程的传统机器学习方法。

近年来,随着深度学习的发展,和机器学习方法相比,深度学习可以自动学习文本特征,具有泛化性更强的优点。因此,研究者提出了许多基于神经网络的方法,深度神经网络被广泛应用于多标签文本分类任务。Berger 等人[11]在词嵌入层,利用预训练模型 word2vec 来捕获单词顺序,将向量输入到 CNN 和 GRU 上,相比传统的词袋模型分类性能得到提升;Baker[12]设计了一种基于 CNN 架构的标签共现的多标签文本分类方法;Liu 等[13]针对极端多标签文本分类中巨大标签空间引发的数据稀疏性和可扩展性,考虑到标签共现问题,提出了 XML-CNN,用卷积神经网络设计了动态池处理文本分类;Chen 等人[14]提出 CNN 与 RNN 的融合机制模型,将 CNN 的输出作为 RNN 的输入,来捕获文本的局部和全局语义信息,再进行多标签的分类任务;Yang 等[15]首次提出将序列生成思想应用于多标签文本分类;Qin 等[16]延续序列生成思想,构建新的训练目标,以便 RNN 能发现最佳标签顺序;Hinton 等[17-18]提出的胶囊网络模型,采用向量神经单元和动态路由更新机制,改进了卷积神经网络模型,克服了 CNN 的弊端,在图像处理领域已取得较好成果;Zhao 等人[19]首次将胶囊网络模型应用在文本分类任务上,其分类效果比 CNN 有一定的提升。

2 BiGRU-Capsule 模型

多标签文本分类问题的目标,是为每个未分类文本样本标注合适的类别标签。可定义为:$X = R^m$ 表示输入样本有 m 维特征空间,$Y = \{y_1, y_2, \cdots, y_n\}$ 表示所有类别标签集合,共有 n 个类别标签。通过训练样本集 $D = \{(x_i, Y_i) \mid 1 \leqslant i \leqslant k\}$,学习得到了一个分类器 $f: X \to 2^Y$。其中,$x_i \in X$ 是输入空间 X 的训练样本,$y_i \in Y$ 是 x_i 的类别标签集合。每个样本都有一个标签集合与之关联,最后通过分类器得到测试样本的所属标签集合[20]。基于 BiGRU-Capsule 的多标签文本分类模型整体结构如图 1 所示。

传统的卷积神经网络在池化层中进行标量计算的操作,在该过程中可能会导致文本特征的丢失。针对上述问题,“神经网络之父”Hinton 提出了胶囊模型,该模型将池化层的标量计算改为向量计算,用神经元向量代替传统卷积神经网络的单个神经元节点,从而确保更多信息不丢失,该神经元向量就是“胶囊”(Capsule)。本文模型是在胶囊模型的基础上构建的,该模型经过 BiGRU 和胶囊模型分别提取到文本的全局和局部特征,最后使用 sigmoid 分类器输出标签。

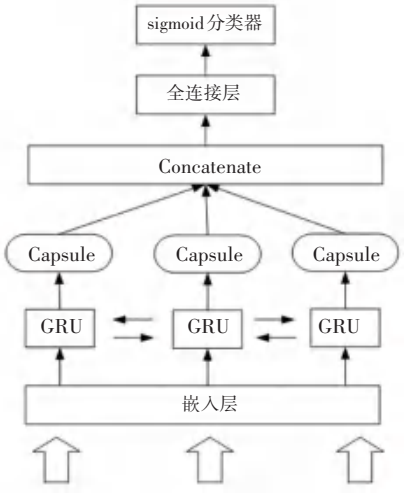

图 1 BiGRU-Capsule 模型整体结构

Fig. 1 Overall structure of BiGRU-Capsule model

2.1 LSTM 模块

循环神经网络(RNN)是一种用于处理序列数据的神经网络。标准的 RNN 结构中有重复的神经网络模块,以链式的形式存在。这个重复模块只有单一神经网络层,因此 RNN 在处理长时间依赖会出现梯度消失和梯度爆炸的问题。为此,Graves 等人[21]提出了长短时记忆网络(Long Short-Term Memory,LSTM)。LSTM 模型在 RNN 的基础上加入了记忆单元和门控机制,使其可以选择性的记住或遗忘信息,从而使时间序列上的记忆信息可控,具备长期记忆功能,在更长的序列中有更好的表现。LSTM 的重复模块有 4 个神经网络层,以一种非常特殊的方式进行信息的交互,每个 LSTM 模块都具有记忆能力。一个 LSTM 重复模块如图 2 所示。

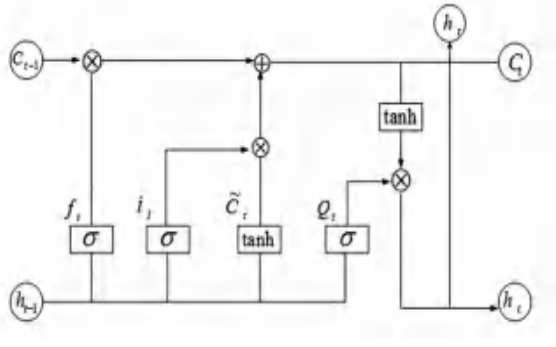

图 2 LSTM 模块

Fig. 2 LSTM module

LSTM 重复模块的神经网络计算公式如下：

$$i_t = \delta(W_i \cdot |h_{t-1}, x_t| + b_i), \quad (1)$$

$$f_t = \delta(W_f \cdot |h_{t-1}, x_t| + b_f), \quad (2)$$

$$o_t = \delta(W_o \cdot |h_{t-1}, x_t| + b_o), \quad (3)$$

$$C_t = \tanh(W_c \cdot |h_{t-1}, x_t| + b_c), \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * C_{t-1}, \quad (5)$$

$$h_t = o_t * \tanh(C_t). \quad (6)$$

其中，i_t、f_t、o_t 分别表示输入门、遗忘门、输出门；x、h、c 表示输入层、隐藏层、记忆单元；$\boldsymbol{W}$ 是权重矩阵；$\boldsymbol{b}$ 是偏置向量。

单向 LSTM 只能提取输入文本上文特征，而不能采集到下文的文本特征，为了能够得到文本全局的语义信息，Schuster 等人[22]提出了双向循环神经网络。将单向网络结构变为双向网络结构，在一定程度上解决了梯度爆炸或梯度消失问题，并且利用当前词的上下文信息提取出输入文本的全局特征表示。将前后 2 个输出向量，得到最终提取的文本特征向量 $\boldsymbol{h}_t$，如式(7)：

$$\boldsymbol{h}_t = [\overrightarrow{h_t} \oplus \overleftarrow{h_t}]. \quad (7)$$

2.2　BiGRU 模块

GRU 可以看做 LSTM 的一种变体，2014 年由 Cho 等人[23]提出。将 LSTM 中隐藏状态和细胞状态合并成一种状态，相比 LSTM 的模型要简单，参数更少、更容易收敛，缩短了训练时间，常用来构建大训练量的模型。GRU 首先读取词嵌入向量 $\boldsymbol{x}_t$ 以及隐藏层状态向量 $\boldsymbol{h}_{t-1}$ 后，经过门控，计算产生输出向量 $\tilde{\boldsymbol{h}}_t$ 和隐藏层状态向量 $\boldsymbol{h}_t$。计算公式如下：

$$z_t = \sigma(W_i * [\boldsymbol{h}_{t-1}, x_t]), \quad (8)$$

$$r_t = \sigma(W_r * [\boldsymbol{h}_{t-1}, x_t]), \quad (9)$$

$$\tilde{\boldsymbol{h}}_t = \tanh(W_c * [r_t \cdot \boldsymbol{h}_{t-1}, x_t]), \quad (10)$$

$$\boldsymbol{h}_t = (1 - z_t) \cdot c_{t-1} + z_t \cdot \tilde{\boldsymbol{h}}_t. \quad (11)$$

其中，σ 是 *sigmoid* 函数；z_t 是一个更新门，控制信息流入下一个时刻；r_t 是一个重置门，控制信息丢失，二者共同决定隐藏状态的输出。本文所使用的 BiGRU 结构如图 3 所示。

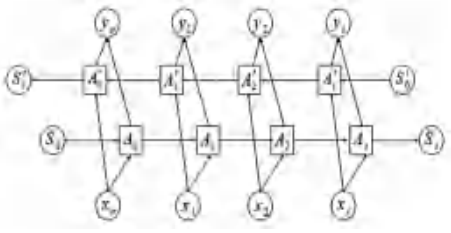

图 3　BiGRU 结构

Fig. 3　BiGRU structure

2.3　Capsule 模块

胶囊网络(CapsNet)的创新，在于提出了输入是向量，输出也是向量的方法。传统的卷积神经网络通过池化层来获取文本的局部特征，但在池化层的操作过程中会造成信息的损失，降低了模型的效率，难以有效地进行编码，且缺乏文本表达能力。胶囊网络中使用神经元向量代替传统神经网络的单个神经元节点，该神经元向量就是所谓的“胶囊”(Capsule)。通过动态路由(Dynamic Routing)训练神经网络，自动的学习单词之间存在的联系，实现向量的信息传递，不仅获取单词在文本中的位置信息，还可以捕获文本的局部空间特征。Capsule 中的激活向量是某个类别特定实体的特征表示，对每个不同的类别，输出不同的向量。向量的模长表示属于该类别的概率，用激活向量的方向表征对应实例的参数。在传统的神经网络中，一般选择 *Sigmoid*、*Relu* 等作为激活函数，但在胶囊网络中提出了新的激活函数 *Squashing*。一个 Capsule 结构如图 4 所示。计算方法如式(12)-式(16)所示。

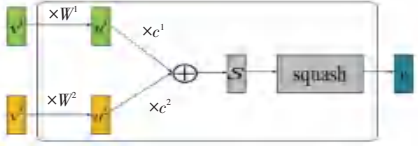

图 4　Capsule 结构

Fig. 4　Capsule structure

$$\hat{u}_{j|i} = W_{ij} \cdot u_i, \quad (12)$$

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ij})}, \quad (13)$$

$$s_j = \sum_i c_{ij} \cdot \hat{u}_{j|i}, \quad (14)$$

$$v_j = \frac{\|s_j\|^2}{0.5 + \|s_j\|^2} \cdot \frac{s_j}{\|s_j\|}, \quad (15)$$

$$b_{ij} \leftarrow b_{ij} + \langle \hat{u}_{j|i}, v_j \rangle. \quad (16)$$

其中，u 为上一层的胶囊输出；c_{ij} 为耦合系数，用来预测上一层胶囊和下一层胶囊之间的相似性；s_j 为 *squashing* 函数的输入；$\boldsymbol{W}$ 为变换矩阵参数；v_j 为输出向量；b_{ij} 的初始值设置为 0 。

3　实验结果及分析

3.1　实验数据集

为了验证模型的有效性，实验采用今日头条 2018 新闻标题多标签语料数据集，该数据集共包含

2 914 000条新闻标题 1 070 个标签。取其中 23 677 条新闻标题作为训练集,5 261 条新闻标题作为测试集。

3.2 实验环境及参数设置

本文的实验环境为 linux 操作系统 openSUSELeap42.3, intel(R) Core(TM) i5-7500 的 CPU, GeForce RTX 2080Ti 的 GPU;编程语言为 python3.6,基于 Keras 框架,TensorFlow 后端实现。为了提高训练的效率,实验模型参数设置为:句子最大长度为 50, *patience* 为 5,*batch_size* 为 32,*dropout* 率为 0.5,*BiGRU* 隐藏层维数为 256,胶囊数量为 32,实验最大迭代次数为 20。

3.3 实验评价指标

在多标签文本分类中,一般采用 *sigmoid* 函数作为输出层的激活函数,使用二分类交叉熵函数(binary_crossentropy,BCE)作为损失函数。即将最后分类层的每个输出节点使用 *sigmoid* 激活函数激活,然后对每个输出节点和对应的标签计算交叉熵损失函数。公式如下:

$$BCE(x)_i = -[y_i \log f_i(x) + (1-y_i)\log(1-f_i(x))], \tag{17}$$

其中, x 为输入; C 为分类类别数; i 属于[1, C]; y_i 为第 i 个类别对于的真实标签。

本文采用的准确率(*accuracy*)是 keras 中的 top_k_categorical_accuracy。准确率的计算公式如下:

$$acc = \frac{T}{T+F+N}, \tag{18}$$

其中, T 代表正确分类的文本数量; F 代表错误分类的文本数量; N 代表属于该类但未被分到该类别的文本数量。*accuracy* 的计算公式,是得到预测对的样本数与总样本数的比值。categorical_accuracy 要求:样本在真值类别上的预测分数,是所有类别预测分数的最大值才算预测正确。不同的是, *accuracy* 针对的是真实标签(y_true)和预测标签(y_pred),都为具体标签的情况,而 categorical_accuracy 针对的是 y_true 为 one-hot 标签, ***y_pred*** 为向量的情况。对于 top_k_categorical_accracy 来说就是计算 *top* - *k* 正确率,当预测值的前 k 个值中存在目标类别即认为预测正确。

假设有 4 个样本,其 *y_true* 为[[0, 0, 1], [0, 1, 0], [0, 1, 0], [1, 0, 0]], *y_pred* 为[[0.1, 0.6, 0.3], [0.2, 0.7, 0.1], [0.3, 0.6, 0.1], [0.9, 0, 0.1]]。则 categorical_accuracy 计算方法为:

(1)将 *y_true* 转为非 onehot 的形式,即 *y_true_new* = [2, 1, 1, 0]。

(2)将 *y_pred* 转为标量标签。其原理是:选取预测向量中最大值所在索引位置作为预测标签,即得到 *y_pred_new* = [1, 1, 1, 0]。

(3)将 *y_true_new* 和 *y_pred_new* 代入公式(18)中计算,得到最终的 *categorical_accuracy* 为 75%。

top_k_categorical_accuracy 的计算方法与 k 息息相关。将 *y_pred* 转为标量标签的原理是:选取预测向量中最大 k 个值所在索引位置作为预测标签,top_k_categorical_accuracy 具体计算方法为:

(1)将 *y_true* 转为非 onehot 的形式,即 *y_true_new* = [2, 1, 1, 0]。

(2)计算 *y_pred* 的 *top_k* 的 label。如 k = 2 时, *y_pred_new* = [[0, 1], [0, 1], [0, 1], [0, 2]]。

(3)根据每个样本的真实标签是否在预测标签的 *top_k* 内,来统计准确率。以上述 4 个样本为例,2 不在[0, 1]内,1 在[0, 1]内,1 在[0, 1]内,0 在[0, 2]内,4 个样本总共预测对了 3 个。因此, k = 2 时 *top_k_categorical_accuracy* = 75%。

本文使用的数据集的标签数有 1 070 类,每个样本的标签数多且不定,因此将 k 设置为 33。(在实验结果与分析中,将 *top_k_categorical_accuracy* 简写为 *tk* - *acc*。)

3.4 实验结果与分析

今日头条 2018 新闻标题多标签语料数据集多标签文本分类结果见表 1。

表 1 多标签文本分类结果

Tab. 1 Multi-label text classification results

Model	*tk* - *acc*
CapsNet	0.726 7
BiGRU-Capsule1(一层 Capsule)	0.761 1
BiGRU-Capsule3(三层 Capsule)	0.791 5

本文实验的主要模型是基于胶囊网络中的 capsule,且在实验过程中发现,叠加多层胶囊能达到较好的效果。因此,设置 CapsNet 模型和一层胶囊模型与本文模型做对比实验。由实验结果可见,本文实验模型效果最好。其准确率随着迭代次数的变化曲线如图 5 所示。BiGRU-Capsule 模型对比 CapsNet 模型和 BiGRU-Capsule1,在准确率上分别提升了 3.48%、3.04%。证明与只能获取局部特征的 CapsNet 模型相比,本文实验模型的效果更佳。CapsNet 模型平均训练一轮的时间约 279 s,BiGRU-Capsule1 模型平均训练一轮的时间约 376 s,BiGRU

-Capsule3 模型平均训练一轮的时间约 110 s，可见本文模型的训练时间比其它两个模型的训练时间更短。实验结果表明，本文实验模型不仅提高了模型的准确率，同时加快了训练速度。

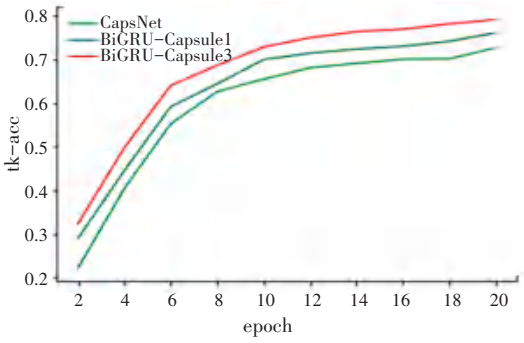

图 5　模型准确率变化曲线图

Fig. 5　The model accuracy rate change curve

4　结束语

本文提出了一种基于 BiGRU-Capsule 的模型，用于多标签文本分类研究。利用 BiGRU 对词的上下文信息提取文本的全局特征，Capsule 有效提取文本的局部特征。实验结果表明，本文提出的模型在 F1 指标上优于对比模型，有效地提升了多标签文本分类的性能。然而，在多标签文本分类领域，仍然有许多问题值得探索，因此下一步工作将研究预训练模型的输出特征与本文设计的模型输出特征进行各种融合操作。

参考文献

[1] TANG D, QIN B, LIU T. Document modeling with gated recurrent neural network for sentiment classification[C]//Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. 2015:1422-1432.

[2] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C]// Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016:1480-1489.

[3] GOPAL S, YANG Y. Multilabel classification with meta-level features[C]//Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010: 315-322.

[4] ELISSEEFF A, WESTON J. A kernel method for multi-labelled classification[C]//Advances in neural information processing systems. 2002: 681-687.

[5] 陆凯，徐华. ML-kNN 算法在大数据集上的高效应用[J]. 计算机工程与应用，2019,55(1):84-88.

[6] BOUTELL M R, LUO J, SHEN X, et al. Learning multi-label scene classification[J]. Pattern recognition, 2004, 37(9): 1757-1771.

[7] TSOUMAKAS G, KATAKIS I. Multi-label classification: An overview[J]. International Journal of Data Warehousing and Mining (IJDWM), 2007, 3(3): 1-13.

[8] READ J, PFAHRINGER B, HOLMES G, et al. Classifier chains for multi-label classification[J]. Machine learning, 2011, 85(3): 333.

[9] OSOJNIK A,PANOV P,DŽEROSKI S. Multi-label classification via multi-target regression on data streams[J]. Machine Learning,2017,106(6):745-770.

[10] 刘慧婷,冷新杨,王利利,等. 联合嵌入式多标签分类算法[J]. 自动化学报,2019,45(10):1969-1982.

[11] BERGER M J. Large scale multi-label text classification with semantic word vectors[R]. Technical report, Stanford University, 2015.

[12] BAKER S,KORHONEN A. Initializing neural networks for hierarchical multi-label text classification[C]// Proceedings of the 2017Conference on Biomedical Natural Language Processing. Strouds-burg, PA: Association for Computational Linguistics, 2017:307-315.

[13] LIU J, CHANG WC, WU Y, YANG Y. Deep learning for extreme multi-label text classification. In: Proc. of the 40th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2017: 115-124.

[14] CHEN G, YE D, XING Z, et al. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization[C]//Neural Networks (IJCNN), 2017 International Joint Conference on. IEEE, 2017: 2377-2383.

[15] YANG P,SUN X,LI W,et al. SGM:sequence generation model for multi-label classification[C] // Proceedings of the 27th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2018:3915-3926.

[16] QIN K,LI C,PAVLU V,et al. Adapting RNN sequence prediction model to multi-label set predictio[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2019:3181-3190.

[17] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules[C]//Advances in neural information processing systems. 2017: 3856-3866.

[18] HINTON G E, SABOUR S, FROSST N. Matrix capsules with EM routing [C]// International Conference on Learning Representations. Toronto, Canada, 2018.

[19] ZHAO W, YE J, YANG M, et al. Investigating capsule networks with dynamic routing for text classification[J]. arXiv preprint arXiv:1804.00538, 2018.

[20] 张丹普. 多标签集成学习算法的关键技术研究[D].北京:中国科学院大学, 2015.

[21] GRAVES A. Supervised Sequence Labelling with Recurrent Neural Networks[J]. Studies in Computational Intelligence, 2012:5-13.

[22] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11):2673-2681.

[23] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer Science, 2014: 1724-1734.